SOME REMARKS ON BAYESIAN PERTURBATION DIAGNOSTICS AND ROBUSTNESS

by

Seymour Geisser
University of Minnesota

Technical Report #521
November 1988

*This work was sponsored in part by an NIH grant GMS-2527

0. Introduction

To paraphrase the opening remarks of G.E.P. Box (1980) "No [Bayesian] analysis can safely be assumed to be adequate. Perspicacious criticism employing diagnostic checks must therefore be applied."

A Bayesian analysis may depend critically on the modeling assumptions which include prior, likelihood and loss function. While a loss function is presumably a choice made in the context of particular situations, there is no harm and potentially some gain in investigating the effect on an analysis using alternative loss functions. The likelihood is supposed to represent to some approximation the physical process generating the data while the prior reflects subjective views about some of the assumed constructs of this process. Now a likelihood model that has been judged adequate in previous situations similar to a current one is certainly a prime candidate for modeling. However even in such situations the statistician is still obliged to investigate its present adequacy. A way of addressing this problem is to perturb the "standard" model to a greater or lesser degree in potentially conceivable directions to determine the effect of such alterations on the analysis. While for the strict Bayesian the prior is subjective, it is common knowledge how difficult it often is to subject an investigator or even a statistician to an elicitation procedure that convincingly yields an appropriately subjective prior. Hence to perturb an investigator's prior or some standard one that appears appropriate, is also sensible.

1. Types of Perturbation

Even when a standard statistical model has proven adequate in data sets

similar to a current one at hand, one is obliged to consider the effect of perturbing the standard model in one way or another on the analysis especially if graphical procedures indicate the possibility that the standard model may only be marginally adequate.

There are a large number of possible perturbation schemata. A typically useful one is where $\omega \epsilon \Omega$ an index governing a perturbation schema is a set of hyperparameters. For $X^{(N)} = (X_1, \ldots, X_N)$, a set of observables, a rather simple example is

$$f(x^{(N)} | \theta, \omega) \propto \prod_{j=1}^{N} \left[1 + \frac{(x_j - \mu)^2}{\omega \sigma^2} \right]^{-(\omega+1)/2}, \qquad \omega \geq 1$$

where the standard is $\omega \to \infty$, i.e. the normal distribution and the most deviant $\omega = 1$, the Cauchy distribution.

A second set is exemplified by a mixture e.g.

$$f(x^{(N)} | \theta, \omega) \propto \prod_{j=1}^{N} [\omega f_1(x_j | \alpha) + (1-\omega) f_2(x_j | \beta)], \quad 0 \leq \omega \leq 1$$

where say $\omega = 1$ is the standard and α and β are subsets of θ.

Use of ω as an indicator is relevant to situations where ω changes the model distribution to varyingly different but known distributional forms not necessarily in the same family. Although this can often be regarded as a special case of either of the first two methods it is best to consider it separately.

A fourth possibility is the use of ω as an exclusion indicator i.e. $X^{(N)} = (X_1, \ldots, X_N)$ has some standard distributional form under ω_0 but for $\omega \neq \omega_0$ one or more of the X_i's have either another distributional form or a completely

unspecifiable distribution. In the former case this could mean for example that an observation's variance differs from the others or more generally that a parameter set not under scrutiny differs for a few of the observations. The latter situation is typically reflected in problems with outliers and aberrant observations that defy satisfactory alternative modeling.

A fifth possibility has to do with what one may term periparametric models. Here $\omega = \omega_0$ specifies a standard density while $\omega \neq \omega_0$ specifies all model densities $f(x^{(N)}|\omega)$ that are within a given neighborhood of $f(x^{(N)}|\omega_0)$ determined by varying ω.

A sixth may have to do with possibly inaccurate measurement of the covariates under $\omega \neq \omega_0$ or even the actual responses themselves. All of the above have to do essentially with perturbation of the likelihood. Similar remarks may be made regarding the prior $g(\theta|\omega)$ and combinations of both likelihood and prior. As a typical example the prior could be a mixture e.g.

$$g(\theta|\omega) = \omega\, g_1(\theta) + (1-\omega) g_2(\theta)$$

with $\omega = 1$ resulting on the standard $g_1(\theta)$ based on previous information while $g_2(\theta)$ expresses the possibility of another view of the situation. This, for whatever it is worth, results in simpler calculations than having to deal with a likelihood mixture. In particular the use of periparametric perturbation models for additional uncertainty about a "standard" prior seems to be a promising approach especially when combined with a standard likelihood. Here one can examine the extent to which bounds on the "standard" prior can be expanded and still yield moderate sample size robustness, e.g. Lavine (1988).

2. Formal Parametric Analyses

A formal Bayesian framework for a perturbation analysis either for a "relevant" parameter or future observables can be delineated.

For the relevant parameter say θ, we can consider the modeling is such that for a given perturbation index ω, the posterior probability function for θ is specified as

$$\mathcal{P}(\theta|x^{(N)},\omega) \propto f(x^{(N)}|\theta,\omega)\ g(\theta|\omega)$$

where $g(\theta|\omega)$ is an assumed prior density for θ conditional on $\omega \in \Omega$ where $\omega = \omega_0$ is the standard.

A loss function

$$L(a,\theta)$$

for taking action $a(X^{(N)}) \in \underset{\sim}{A}$ upon observing $X^{(N)}$, given θ is the true value, is assumed (the loss function itself may also be perturbed but we shall not consider this possiblity in what follows). The average loss

$$\bar{L}_\omega(a) = \int L(a,\theta)\ \mathcal{P}(\theta|x^{(N)},\omega)\ d\theta,$$

which depends on ω, is now minimized

$$\min_a \bar{L}_\omega(a) = \bar{L}_\omega\ (a^*_\omega)$$

yielding optimal action a^*_ω when ω is "true". We then consider the difference in the loss when taking action $a^*_{\omega_0} = a^*$, the optimal action under the standard and when $\omega \neq \omega_0$ is true. We define the differential loss as

$$d(\omega) = \bar{L}_\omega(a^*) - \bar{L}_\omega(a^*_\omega) \geq 0.$$

One then can examine this loss over a possible range of ω to assess its importance with regard to the action taken under ω_0 and in particular $d^* = \max_{\omega\in\Omega} d(\omega)$. We could also assess its local significance by examining $d(\omega)$ in a neighborhood about ω_0. In fact if ω is a scalar and the second derivative of $d(\omega)$ exists and is continuous the calculation of the curvature at $\omega = \omega_0$ i.e. $d''(\omega_0)$, since $d'(\omega_0) = 0$, could be rather informative regarding local perturbations. For example a large curvature would indicate that the actions taken could be highly sensitive to a slight variation in the standard model.

For a vector ω, the matrix of second derivatives will govern the local curvature and one could assess the maximum curvature i.e. in the direction of the normed vector associated with the largest root of the matrix of second derivatives evaluated at the standard $\omega = \omega_0$. Cook (1986) has proposed probing local curvature with regard to the displacement of maximized log-likelihoods.

3. Predictive Analysis

We now outline the situation for prediction. The model considered is the joint probability function

$$f(x^{(N)}, x_{(M)}, \theta | \omega) = f(x_{(M)} | x^{(N)}, \theta, \omega)\ f(x^{(N)} | \theta, \omega)\ g(\theta | \omega)$$

whence we obtain

$$f(x_{(M)} | x^{(N)}, \omega) = \frac{\int f(x^{(N)}, x_{(M)}, \theta | \omega)\ d\theta}{\int\int f(x^{(N)}, x_{(M)}, \theta | \omega)\ d\theta\ dx_{(M)}}$$

Now assume that $L(a, x_{(M)})$ is the loss incurred in taking action a when observing $x^{(N)}$ given a future realization $x_{(M)}$. The average predictive loss

$$\bar{L}_\omega(a) = \int L(a, x_{(M)})\ f(x_{(M)} | x^{(N)}, \omega)\ dx_{(M)}$$

is then minimized

$$\min_a \bar{L}_\omega(a) = \bar{L}_\omega(a^*_\omega)$$

where a^*_ω is the optimal action. As before letting $a^*_{\omega_0} = a^*$, we define the differential loss as $d(\omega) = \bar{L}_\omega(a^*) - \bar{L}_\omega(a^*_\omega)$ and examine globally $\max_\omega d(\omega)$ to determine the possible extent of the maximum effect of the perturbations.

Further in regular cases one can again study locally the maximum curvature which occurs in the direction of the normal vector associated with the largest root of the Hessian matrix, say $d''(\omega_0)$. If local curvature is appreciable it would appear that the sample is not even robust locally and a review of the standard model is in order. Of course if the perturbed ω model is deemed

reasonable one possibility is to define a prior distribution for ω and then integrate it out to obtain

$$f(x_{(M)}|x^{(N)}) = \int g(\omega)\ f(x_{(M)}|x^{(N)},\omega)\ d\omega.$$

4. Other Perturbation Diagnostics

Often, we are not in a position to discuss decisions or actions which would necessarily flow from a data set and consequently report either the posterior or predictive distribution itself or some high probability density region for θ or $X_{(M)}$. For reporting the entire posterior distribution the Kullback-Leibler estimative divergence,

$$K(\mathcal{P}_{\omega},\mathcal{P}_{\omega_o}) = E[\log \mathcal{P}_{\omega} - \log \mathcal{P}_{\omega_o}],$$

where $\mathcal{P}_{\omega} = \mathcal{P}(\theta|y^{(N)},\omega)$, is a reasonable diagnostic to consider when it exists and is finite, Geisser (1985), Johnson and Geisser (1985), McCulloch (1986) and can be investigated in a variety of paradigms. Similarly for predictive distributions a predictive divergence

$$K(\omega,\omega_o) = E[\log f_{\omega} - \log f_{\omega_o}]$$

where $f_{\omega} = f(x_{(M)}|x^{(N)},\omega)$, will serve as a reasonble diagnostic. Divergences of this sort were introduced by Johnson and Geisser (1982,1983) for determining influential observations, one of the particular types of perturbation previously mentioned, and were termed predictive influence functions (PIF).

Both, estimative and predictive diagnostics, are most useful in indicating the relative effect of various perturbations.

There may be, however, some difficulty in adequately interpreting globally

$$\max_{\omega\in\Omega} K(\mathcal{P}_{\omega}, \mathcal{P}_{\omega_o}), \text{ or } \max_{\omega\in\Omega} K(\omega,\omega_o)$$

for some of these paradigms.

Another use is to find the direction in which local perturbations have the greatest effect in terms of normal curvature. It can be shown that under suitable regularity conditions that the matrix of second derivatives of $K(\mathcal{P}_{\omega},\mathcal{P}_{\omega_o})$ or $K(\omega, \omega_o)$ for ω a vector of perturbations, say

$$K''_{\omega=\omega_o} = I(\omega_o)$$

where $I(\omega_o)$ is the Fisher Information matrix for either the posterior or predictive distribution at $\omega = \omega_o$, Kullback (1959). The curvature in direction z where $\omega(t) = \omega_o + tz$ and $z'z = 1$ is

$$C_z = z'I(\omega_o)z$$

so that the maximum curvature C^* is in the direction z^*, the vector associated with the maximum root of $I(\omega_o)$, where

$$C^* = z^*I(\omega_o)z^*.$$

An examination of the components of z^* will indicate which ones, namely the larger ones, are those perturbations which relatively most alter the posterior or predictive distribution in terms of the divergence.

Once potentially significant directions are identified, an analysis involving these directions is in order to ascertain whether local departures for them are important enough to vitiate the standard analysis.

The L^1 norm between two densities f and g, favored by Devroye (1987), or the L^2 norm between $\sqrt{f}$ and $\sqrt{g}$ favored by Pitman (1979) as measures of distance between densities can also be used here as diagnostics for the posterior distribution. More generally the Hellinger distance between densities raised to the n-th power

$$H^n = \int |f^{1/n} - g^{1/n}|^n dx$$

yields these as special cases. For the case here with n=2 we have

$$H^2(\wp_\omega, \wp_{\omega_o}) = \int(\wp_\omega^{1/2} - \wp_{\omega_o}^{1/2})^2 \, d\theta$$

for posterior densities and for predictive densities,

$$H^2(\omega, \omega_o) = \int(f_\omega^{1/2} - f_{\omega_o}^{1/2})^2 \, dx_{(M)} ,$$

which accords with the norm favored by Pitman. Under suitable smoothness conditions, twice the matrix of second derivatives evaluated at $\omega=\omega_o$ when ω is a vector, is

$$2H^{2''}_{\omega=\omega_o} = I(\omega_o)$$

again Fisher's Information matrix, which is an indicator of the usefulness of this quantity for local perturbation analysis.

The L^1 norm may also be used. While it is unaffected by any one-to-one transformation as is the divergence and H^2, it is analytically awkward and does not discriminate between differences of the two densities when the smaller of the two is large or small, as does H^2 and the divergence.

While the divergence and H^2 are as sensible as any measure of how densities differ overall it is difficult to establish a reasonable calibration that different values of the divergence or H^2 entail except in a relative sense. Methods for a more suitably direct interpretation that a statistician, and more to the point an investigator, can readily understand can also be defined but they involve rather specific situations. We now present some of these ways of assessing the robustness in terms of posterior or predictive regions for θ or $X_{(M)}$. One could restrict oneself to perturbations that could matter as determined locally but we shall retain the same notation as before for two reasons. First for convenience in that it is possible that the entire ω set may matter and secondly in certain instances one may not be specifically interested

in a local determination. The potential value of the local analysis is the possibility of restricting the dimension of the vector of perturbations to a small set that can more easily be managed by the assessments we now shall propose.

The first method is to assess the robustness of a $1-\alpha$ highest probability density region based on the standard ω_o. Suppose this region denoted by $R_{1-\alpha}(\omega_o)$ has volume $V(\omega_o)$ and when perturbed the highest probability density region $R_{1-\alpha}(\omega)$ has volume $V(\omega)$. Let $\nu(\omega)$ be the volume of the intersection of $R_{1-\alpha}(\omega)$ and $R_{1-\alpha}(\omega_o)$ as a function of ω,

$$\nu(\omega) = \text{volume } [R_{1-\alpha}(\omega) \cap R_{1-\alpha}(\omega_o)]$$

and let

$$\Gamma_\omega = \frac{\nu(\omega)}{M(\omega)} \quad ,$$

where $M(\omega) = \max (V(\omega), V(\omega_0))$, be the ratio of the volume of the intersection to whichever is larger the standard or the perturbed for the given ω. Then calculate

$$\min_{\omega \in \Omega} \Gamma_\omega = \Gamma_{\omega^*}$$

which now yields the proportion of the region for the "worst" possible case at a given probability $1-\alpha$. Hence one has an easily interpretable value for assessing the robustness of the data set in terms of a standard analysis involving a $1-\alpha$ region in the presence of presumably anticipated perturbations. A second method focuses on the use of the standard region's $R_{1-\alpha}(\omega_o)$ perturbed probability when $\omega \neq \omega_0$. Here we use either

$$\Pr[\theta \epsilon R_{1-\alpha}(\omega_o)|\omega] = \int_{R_{1-\alpha}(\omega_o)} \wp(\theta|y^{(N)},\omega)d\theta = 1-\alpha_\omega$$

or

$$\Pr[X_{(M)} \epsilon R_{1-\alpha}(\omega_o)|\omega] = \int_{R_{1-\alpha}(\omega_o)} f(x_{(M)}|x^{(N)},\omega)\, dx_{(M)} = 1-\alpha_\omega$$

and either

$$\max_{\omega\epsilon\Omega} |\, 1-\alpha-(1-\alpha_\omega)| = \max_{\omega\epsilon\Omega} |\, \alpha_\omega - \alpha|$$

or

$$\max_{\omega\epsilon\Omega} \frac{|\alpha_\omega-\alpha|}{1-\alpha_\omega}$$

as easily interpretable values. This second method is most compelling when some specified region is critical to an analysis, e.g. the effect of the perturbation on the calculation of the probability of an observable exceeding some threshold.

In fact as a very simple illustration of this consider $X^{(N)} = (X_1,\ldots,X_N)$ a random sample from

$$f(x|\theta,\omega) = \theta\ e^{-\theta(x-\omega)} \qquad x \geq \omega \geq 0$$

and noninformative prior

$$g(\theta) \propto \theta^{-1}\ .$$

Suppose $x_1,\ldots,x_d$ are fully observed realizations and $X_{d+1},\ldots,X_N$ are independently censored at values $x_{d+1},\ldots,x_N$. We further suppose, as is almost always the case, that

$$x_{(1)} = \min\,(x_1,\ldots,x_d) \leq \min\,(x_{d+1},\ldots,x_N)\ .$$

The predictive distribution function, Geisser (1982), is then easily calculated to be

$$\Pr\,[X_{N+1} \leq x|\ x^{(N)},\ \omega] = 1 - \left(1 + \frac{x-\omega}{N\bar{x}}\right)^{-d} \qquad \begin{matrix} 0 \leq \omega \leq x_{(1)}. \\ \omega \leq x \end{matrix}$$

Here it is of interest to calculate the probability of a survival threshold

say y

$$\Pr\left[X_{N+1} > y \mid X^{(N)}, \omega\right] = \left(1 + \frac{y - \omega}{N\bar{x}}\right)^{-d}$$

where the standard say is $\omega=0$. Of course the divergence and the H^n distances are largely irrelevant for this case but we can easily calculate

$$\max_{0\le\omega\le x_{(1)}} \left|\alpha_{\omega}(y) - \alpha(y)\right| = \left(1 + \frac{y - x_{(1)}}{N\bar{x}}\right)^{-d} - \left(1 + \frac{y}{N\bar{x}}\right)^{-d}$$

for a fixed y or conversely for those values of y such that the quantity on the right is no larger that a given value considered negligible with respect to stating a probability for surviving the threshold.

As an example consider the following data reported in Gnedenko et al (1969, p. 176) consisting of a sample of N=100 items tested and time to failure recorded for each item until 500 standard time units have elapsed. The recorded failure times for 11 items were: 31, 49, 90, 135, 161, 249, 323, 353, 383, 436, 477. The remaining 89 items survived the test termination time. If interest is focused on the probability of a future item surviving 500 time units then

$$\Pr\left[X_{N+1} > 500 \mid \omega\right] = \left(1 + \frac{500-\omega}{47{,}187}\right)^{-11}$$

$$= .891 \qquad \text{for } \omega = 0$$

$$= .897 \qquad \text{for } \omega = x_{(1)} = 31$$

Hence

$$\max_{0\le\omega\le 31} \left|\alpha_{31}(500) - \alpha(500)\right| = .006$$

On the other hand one might be interested in that value y such that

$$\Pr[X_{N+1} > y \mid \omega] = .5$$

Here for $\omega=0$, $y = 3069$ and for $\omega = 31$, $y = 3100$ yielding a maximum relative difference of 1%. In passing we also point out here that the maxima for the

divergence and the two norms are $K = \infty$, $H^2 \doteq H^1 = .01$ and are not particularly informative. The divergence indicates only a difference in support while the norms are approximately and exactly twice the probability assigned to the largest interval over which only one of the densities is supported.

More generally, implementation of these methods in other cases could involve the algebraic or numerical calculation of the intersection of two n-dimensional hyperellipsoids which could be quite burdensome for $n > 3$.

Even more complex situations arise where the highest probability density regions are disconnected. Here one may also want to take into consideration the distance from the standard a perturbed and disconnected region is in ordering the diagnostics discussed above, i.e. not only the size of the non-intersecting disconnected region but its distance in some sense from the standard.

5. Acknowledgment

This work was sponsored in part by an NIH grant GMS-25271.

References

Box, G.E.P.(1981). Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society A, 143, 383-430.

Cook, R.D.(1986). Assessment of local influence (with discussion). Journal of the Royal Statistical Society B, 48, 2, 133-169.

Devroye, L. (1987). A Course in Density Estimation. Birkhauser.

Geisser, S. (1982) Aspects of predictive and estimative approaches in the determination of probabilities, Biometrics Supplement: Current Topics in Biostatistics and Epidemiology 38, 1, March, 75-85.

Geisser, S. (1985) On the predicting of observables: a selective update, in: Bernardo, J.M. et al. (Ed.) Bayesian Statistics 2, (with discussion) 203-230. Amsterdam, North-Holland.

Gnedenko, B.B., Belyayev, Y.K., and Solovyev, A.D. (1969). Mathematical Methods of Reliability Theory. New York and London: Academic Press.

Johnson, W. & Geisser S. (1982) Assessing the predictive influence of observations, in: G. Kallianpur, P.R. Krishnaiah & J.K. Ghosh (Eds) Statistics and Probability Essays in Honor of C.R. Rao, 343-358. Amsterdam, North-Holland.

Johnson, W. & Geisser, S. (1983) A predictive view of the detection and characterization of influential observations in regression analysis, Journal of American Statistical Association 78, 137-144.

Johnson, W. & Geisser, S. (1985) Estimative influence measures for the multivariate general linear model, Journal of Statistical Planning and Inference 11, 33-56.

Kullback, S. (1959). Information theory and Statistics. New York, John Wiley and Sons.

Lavine, M. (1987). Prior influence in Bayesian Statistics. University of Minnesota Technical Report No. 504

McCulloch, R. (1986). Local prior influence. University of Minnesota Technical Report No 477.

Pitman, E.J.G. (1979). Some Basic Theory for Statistical Influence. London, Chapman and Hall.